Analysing Self-Harm Representations in Language Models: a Cross-Architecture Study

Luis Espinosa-Anke **Carla Perez-Almendros**
Cardiff University
{espinosa-ankel, perezalmendrosc}@cardiff.ac.uk

Abstract

Detecting self-harm is a challenging and high-stakes task requiring the highest accuracy to enable timely intervention and flagging at-risk users. In this paper, we present an analysis of how LLMs represent such self-harm posts.We perform two experiments: (1) We train and evaluate linear probes across all layers of each model on two self-harm datasets and find that self-harm information crystallizes in the final 3 - 7% of network layers (93 to 97% depth). (2) We extract contrastive self-harm directions and, after performing a normalization step, we find that the most accurate probes are not necessarily the most linearly separable. In particular, we find Gemma-3-4B to represent this *contrastive self-harm direction* in a slightly different, more intricate way than the other LLMs [1].

1 Introduction

Self-harm content on social media poses a significant public health concern. Researchers have focused on using AI for analyzing mental health disorders (Owen et al., 2024), with many building classifiers for depression (Nadeem, 2016) and self-harm detection (Antypas et al., 2025; Rozova et al., 2022; Ayre et al., 2021). Automated detection systems, thus, play a critical role in content moderation (Coppersmith et al., 2018; Ji et al., 2022) and for flagging depressive disorders (Leiva and Freire, 2017). Past approaches include using LSTMs (Yates et al., 2017), BERT-based transformers (Owen et al., 2020) and LLMs (Ohse et al., 2024). However, there is still a significant gap in understanding *how* LLMs internally represent such content: where in the network self-harm concepts crystallize, how geometrically stable these representations are across layers, or whether the structure is consistent across architectures and scales. We argue that gaining deeper understanding on these questions can contribute dramatically to building safer systems that go beyond prompt engineering, and instead include explicit, testable components, which also enables more realistic intervention, governance and escalation (Reddy and Reddy, 2025)[2]. In this context, "representation engineering" (Zou et al., 2023) offers a promising framework for understanding model internals through the lens of residual stream activations. In particular, the extraction of contrastive directions (typically via datasets of contrastive prompts that elicit opposed behaviours in LLMs) allows us to probe concept separability as well as to characterise the geometry of safety-relevant embeddings (Turner et al., 2023; Li et al., 2024), with direct applications in bias mitigation (Siddique et al., 2026). In this paper, thus, we ask ourselves the question whether self-harm *content* is equally, more or less likely to be captured by a single direction, a proven feature in "refusal", for instance, a well-known proxy for embedded safety in AI-assistants (Arditi et al., 2024). We found consistent late-layer crystallisation of self-harm concepts, and a cross-layer block-diagonal embedding of contrastive directions, which we argue paves the way for future more effecrtive intervention.

2 Related Work

Automated detection of suicidal ideation and self-harm has long been studied (Leiva and Freire, 2017; Coppersmith et al., 2018; Ji et al., 2022; Yang et al., 2023; Ghosh et al., 2025). However, they are noteably unable to handle mental health-specific challenges like prioritising crisis intervention accuracy or preventing escalations (Zhang et al., 2025; Lee

[1] Code available at https://github.com/luisespinosaanke/self-harm-analysis.

[2] When deployed without such features, AI-based assistants have proven brittle in real world scenarios, and caused genuine public health issues: https://www.bbc.co.uk/news/world-us-canada-65771872.

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

et al., 2025; Schoene and Canca, 2025; Stamatis et al., 2026; Lyu et al., 2025).Seeking AI-safety, LLMs are reinforced with human feedback to replicate human preferences (Ziegler et al., 2019; Bai et al., 2022a). Initiatives like Constitutional AI (Bai et al., 2022b) also allow models to learn from their own outputs to maximize safe and accurate responses. Another approach to safety involves the better understanding of LLMs behaviour through representation analysis, in particular leveraging linear probes (Alain and Bengio, 2016; Belinkov, 2022; Burns et al., 2023). Recent work has extended probing to latent knowledge discovery (Burns et al., 2023) and safety-relevant concepts. Zou et al. (2023) characterizes *where* in the activation space the concept lives and how it evolves across layers, by extracting mean-difference contrastive directions. Subramani et al. (2022) and Turner et al. (2023) showed that such directions can be used for inference-time control, while Siddique et al. (2026) and Li et al. (2024) demonstrated downstream applications in bias mitigation and truthfulness.

3 Data and Models

We evaluate on two datasets. First, **X-Sensitive (X-S)** Antypas et al. (2025), a multi-label corpus of English-language social media posts annotated for six sensitive content categories: self-harm, conflictual, profanity, sex, drugs, and spam. We combine all available splits (train, validation, test) and extract 200 self-harm positive posts (`selfharm`$= 1$) and the 200 control posts where all categories are 0. Second, **SH-Detection (SH-D).** Tharsi (2025), where self-harm posts have `class`$=$`"self-harm"`; all remaining posts serve as controls. Unlike X-S, controls may include posts from any non-self-harm category. We sample a balanced binary 600-post dataset.For both datasets, posts shorter than 20 characters are discarded, URLs are removed, whitespace is collapsed, and @mentions are anonymised. Representative examples are shown in Table 1. In terms of LLMs, we select four instruction-tuned transformer architectures: **Qwen3-0.6B** (0.6B parameters, 28 layers, $d_{\text{model}} = 1024$; Qwen-Team, 2025), **Llama 3.2-1B-Instruct** (1B parameters, 16 layers, $d_{\text{model}} = 2048$; Grattafiori et al., 2024), **Llama 3.2-3B-Instruct** (3B parameters, 28 layers, $d_{\text{model}} = 3072$; Grattafiori et al., 2024), and **Gemma-3-4B-it** (4B parameters, 34 layers, $d_{\text{model}} = 2560$; Gemma-Team, 2025). We access internal representations via TransformerLens (Nanda and Bloom, 2022), hooking into `hook_resid_post` at each transformer block to extract residual stream activations after the full sublayer stack (attention + MLP). We use the *last token position* as the representation of each input, following standard practice for causal LMs where the final token aggregates full left context.

Source	Example post
X-S/S-H	*"feeling suicidal retweet and like to stop me, comment to encourage me help"*
X-S/C	*"Just watched the game, incredible performance tonight"*
SH-D/S-H	*"i'm useless and a waste of space but at least i finally got the courage to off myself"*
SH-D/C	*"done with the stress"*

Table 1: Representative self-harm (S-H) and control (C) examples from X-Sensitive (X-S) and self-harm detection (SH-D).

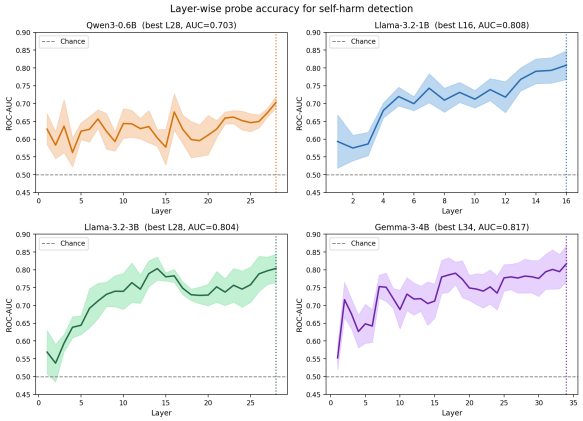

Figure 1: Layer-wise probe ROC-AUC for all four models on X-Sensitive. Shaded regions denote ± 1 SD across 5-fold CV.

4 Experiments

4.1 Layer-wise Probing

Setup. For each layer $\ell \in \{1, \ldots, L\}$, we extract last-token residual-stream activations $\mathbf{a}_\ell \in \mathbb{R}^{d_{\text{model}}}$ for all posts in each dataset (400 for X-S, 600 for SH-D) and train an ℓ_2-regularised logistic regression probe ($C = 1.0$, `lbfgs` solver, max 1000 iterations) using 5-fold stratified cross-validation. We report mean ROC-AUC as our primary metric, as it measures discriminability independently of classification threshold and is robust to class balance effects. Activations are not normalised prior to probing.

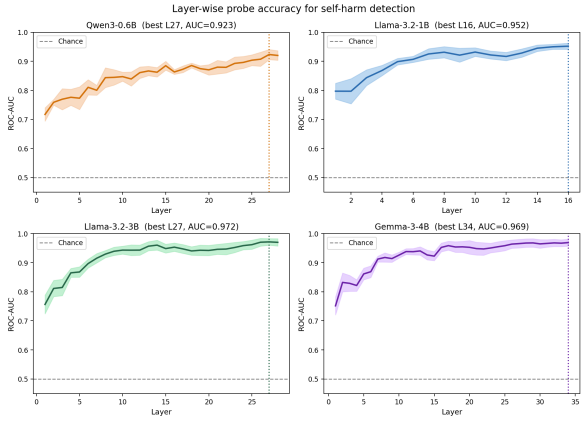

Figure 2: Layer-wise probe ROC-AUC for all four models on SH-Detection. Shaded regions denote ±1 SD across 5-fold CV.

		X-Sensitive		SH-Detection	
Model	L	**Best** ℓ	**AUC**	**Best** ℓ	**AUC**
Qwen3-0.6B	28	28	.703 ± .016	27	.923 ± .017
Llama 3.2-1B	16	16	.808 ± .040	16	.952 ± .011
Llama 3.2-3B	28	28	.804 ± .040	27	.972 ± .011
Gemma-3-4B	34	34	.817 ± .050	34	.969 ± .012

Table 2: Best linear probe ROC-AUC per dataset. L = total layers; Best ℓ is the layer maximising mean ROC-AUC across 5-fold CV.

Results. Table 2 summarises the probing results. Across all four models and both datasets, self-harm information crystallizes at 93 - 97% of network depth. In terms of classification, AUC ranges from 0.703 (Qwen3-0.6B) to 0.817 (Gemma-3-4B) in X-S. On SH-D, all models improve substantially: AUC ranges from 0.923 (Qwen3-0.6B) to 0.972 (Llama 3.2-3B), with significantly lower variance (standard deviation halves). Figures 1 and 2 show the full layer-wise increasing AUC curves for X-S and SH-Detection, respectively. As for a dataset comparison, SH-D achieves uniformly higher AUC with lower variance. We attribute this to corpus differences, e.g., presence of figurative or metaphorical content in X-S's vs. SH-D's cleaner binary labelling. Such figurative language also makes X-S's negative examples subtler, which might reduce probe ceiling performance.

Error analysis. At the best probe layer on X-S, false negatives mostly consist of self-harm posts using figurative or metaphorical language, legal or news contexts, or pop-culture references. False positives, on the other hand, are predominantly mental health advocacy posts, content-warning-prefixed posts, and third-person crisis descriptions, which use vocabulary related to critical situations, but do not convey first-person distress signals, which makes them legitimate confounders.

	X-Sensitive		SH-Detection	
Model	**Best** ℓ	**Cohen's** d	**Best** ℓ	**Cohen's** d
Qwen3-0.6B	28	0.95	27	2.03
Llama 3.2-1B	16	1.26	16	2.06
Llama 3.2-3B	28	1.14	27	2.08
Gemma-3-4B	34	0.71	34	1.18

Table 3: Contrastive direction Cohen's d at each model's best probe layer (from Table 2).

4.2 Contrastive Direction Analysis

Setup. Following Zou et al. (2023), we compute the contrastive self-harm direction at each layer:

$$\mathbf{d}_\ell = \frac{\bar{\mathbf{a}}_\ell^+ - \bar{\mathbf{a}}_\ell^-}{\|\bar{\mathbf{a}}_\ell^+ - \bar{\mathbf{a}}_\ell^-\|} \quad (1)$$

where $\bar{\mathbf{a}}_\ell^+$ and $\bar{\mathbf{a}}_\ell^-$ are mean last-token activations of self-harm and control posts. Each example's scalar projection $s_i = \mathbf{a}_{\ell,i} \cdot \mathbf{d}_\ell$ provides a single-dimensional self-harm score. To enable cross-model comparison despite differences in activation magnitudes, we compute Cohen's d at the best-separation layer:

$$d = \frac{\bar{s}^+ - \bar{s}^-}{\sqrt{(\sigma_+^2 + \sigma_-^2)/2}} \quad (2)$$

We assess directional stability via the $L \times L$ pairwise cosine similarity matrix $\cos(\mathbf{d}_i, \mathbf{d}_j)$.

Separability. Table 3 shows contrastive geometry results. The reason for normalizing with Cohen's d is that raw separation scores vary by orders of magnitude across models and datasets (e.g. 3.1 for Llama-1B on X-S vs. 6,767 for Gemma-3-4B on SH-D) and so comparison between models would be challenging.

On X-S, effect sizes range from 0.71 (Gemma-3-4B) to 1.26 (Llama 3.2-1B), all indicating large-effect separability. On SH-D, effect sizes roughly double: 1.18 to 2.08, with Gemma-3-4B again the lowest despite its superior probe AUC. This discrepancy, namely the high classification accuracy vs. a lower geometric separation for Gemma is consistent for both datasets. We argue that this points to a model-specific representational property rather than an artifact in the dataset. To give a concrete example, on SH-D the contrastive direction for Llama 3.2-3B peaks in Cohen's d at

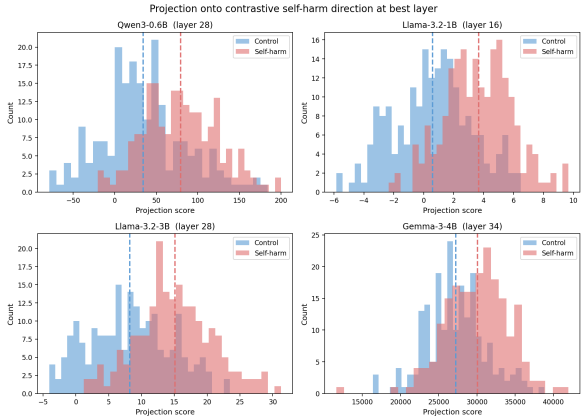

Figure 3: Scalar projections onto the contrastive self-harm direction at each model's best layer for X-S.

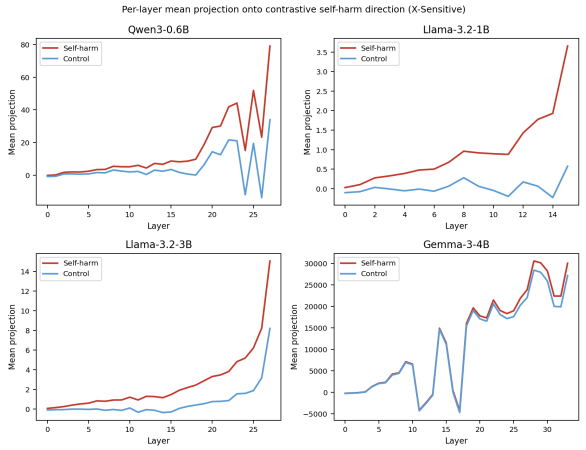

Figure 4: Per-layer mean projection onto the contrastive self-harm direction for all four models (X-Sensitive).

layer 14 ($d = 2.20$), well before the final layers, while probe AUC continues rising to layer 27. This divergence sharpens the interpretation that probing and directional analysis capture complementary aspects: mid-network directions may be potent but not yet sufficient for accurate linear classification.

One geometric interpretation of Gemma's dissociation is as follows. A linear probe has access to the full d_{model}-dimensional activation space and can find a separating hyperplane exploiting any combination of features, including signal distributed, even if weakly, across many dimensions. Cohen's d on the mean-difference direction reveals whether the *mean class centroids* are far apart relative to within-class spread along that single axis. High AUC with low d therefore implies that Gemma's self-harm signal is not concentrated in one direction, but spread across the activation space. Next, we generate projection histograms at each model's best layer (Figure 3), as well as per-layer mean projection curves (Figure 4) for X-S[3]. These visuals show, first, a clear bimodal distribution with overlapping tails, which illustrates the geometric difference between the encodings of this dataset across models (although smaller for Gemma-3-4B). Also, that control means diverge mostly in the final two to four layers.

	X-Sensitive		SH-Detection	
Model	**Adj.**	**Cross**	**Adj.**	**Cross**
Qwen3-0.6B	0.894	0.289	0.908	0.376
Llama 3.2-1B	0.758	0.222	0.801	0.273
Llama 3.2-3B	0.858	0.218	0.879	0.267
Gemma-3-4B	0.897	0.239	0.918	0.135

Table 4: Directional stability summary. **Adj.**: mean cosine similarity between adjacent-layer direction pairs. **Cross**: mean cosine between directions in the first and last thirds of the network. Full $L \times L$ matrices are in Figure 5 (Appendix A).

Directional stability. Table 4 summarises the block-diagonal structure visible in the full cosine matrices (Figure 5, Appendix A). On both corpora, adjacent-layer directions are highly aligned (0.76 to 0.92), while directions from the first and last thirds of each network are nearly orthogonal (between 0.14 and 0.38). This means that the self-harm concept is dynamically *re-represented* across depth, i.e., the direction *rotates* substantially between early and late layers, and this structural property holds on the two datasets. Gemma-3-4B shows the lowest cross-comparison cosine on SH-D (0.135).

5 Conclusion

We have presented a cross-architecture analysis of self-harm representations. In our experiments, self-harm concepts crystallize consistently in the 3-7% "latest" layer block. Moreover, the self-harm contrastive direction is not stable across network depth, a direction extracted at an early layer does not transfer to late layers. In addition, the most accurate linear probes are not the most geometrically separable. Finally, the uniform performance gap between X-Sensitive and SH-Detection (0.12 to 0.17 AUC, $\sim$1 Cohen's d unit) also suggests that labelling quality and corpus composition have a large, systematic effect on representational quality. For future work, we could extend these experiments to larger (7B up to or beyond 70B) models and to other sensitive content categories.

[3] Similar pattern for SH-D.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Dimosthenis Antypas, Indira Sen, Carla Perez Almendros, Jose Camacho-Collados, and Francesco Barbieri. 2025. Sensitive content classification in social media: A holistic resource and evaluation. In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*, pages 17–31, Vienna, Austria. Association for Computational Linguistics.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction.

Karyn Ayre, André Bittar, Joyce Kam, Somain Verma, Louise M Howard, and Rina Dutta. 2021. Developing a natural language processing tool to identify perinatal self-harm in electronic healthcare records. *PLoS one*, 16(8):e0253809.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. In *Biomedical Informatics Insights*, volume 10.

Gemma-Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Soumitra Ghosh, Gopendra Vikram Singh, Shambhavi, Sabarna Choudhury, and Asif Ekbal. 2025. Just a scratch: Enhancing LLM capabilities for self-harm detection through intent differentiation and emoji interpretation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27428–27445, Vienna, Austria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2022. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 9(1):214–226.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering.

Víctor Leiva and Ana Freire. 2017. Towards suicide prevention: Early detection of depression on social media. In *Internet Science - 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings*, Lecture Notes in Computer Science, pages 428–436. Springer.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Chenhan Lyu, Yutong Song, Pengfei Zhang, and Amir M. Rahmani. 2025. Domain-specific constitutional ai: Enhancing safety in llm-powered mental health chatbots. In *2025 IEEE 21st International Conference on Body Sensor Networks (BSN)*, page 1–4. IEEE.

Moin Nadeem. 2016. Identifying depression on twitter.

Neel Nanda and Joseph Bloom. 2022. TransformerLens. `https://github.com/TransformerLensOrg/TransformerLens`.

Julia Ohse, Bakir Hadžić, Parvez Mohammed, Nicolina Peperkorn, Michael Danner, Akihiro Yorita, Naoyuki Kubota, Matthias Rätsch, and Youssef Shiban. 2024. Zero-shot strike: Testing the generalisation capabilities of out-of-the-box llm models for depression detection. *Computer Speech & Language*, 88:101663.

David Owen, Jose Camacho-Collados, and Luis Espinosa Anke. 2020. Towards preemptive detection of depression and anxiety in Twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 82–89, Barcelona, Spain (Online). Association for Computational Linguistics.

David Owen, Amy J Lynham, Sophie E Smart, Antonio F Pardiñas, and Jose Camacho Collados. 2024. Ai for analyzing mental health disorders among social media users: Quarter-century narrative review of progress and challenges. *Journal of medical Internet research*, 26:e59225.

Qwen-Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Pavan Reddy and Nithin Reddy. 2025. Preventing another tessa: Modular safety middleware for health-adjacent ai assistants.

Vlada Rozova, Katrina Witt, Jo Robinson, Yan Li, and Karin Verspoor. 2022. Detection of self-harm and suicidal ideation in emergency department triage notes. *Journal of the American Medical Informatics Association*, 29(3):472–480.

Annika Marie Schoene and Cansu Canca. 2025. 'for argument's sake, show me how to harm myself!': Jailbreaking llms in suicide and self-harm contexts. In *2025 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7.

Zara Siddique, Irtaza Khalid, Liam Turner, and Luis Espinosa-Anke. 2026. Shifting perspectives: Steering vectors for robust bias mitigation in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 809–820, Rabat, Morocco. Association for Computational Linguistics.

Caitlin A. Stamatis, Jonah Meyerhoff, Richard Zhang, Olivier Tieleman, Matteo Malgaroli, and Thomas D. Hull. 2026. Beyond simulations: What 20,000 real conversations reveal about mental health ai safety.

Nishant Subramani, Nivedita Suber, Srinath Mathew, and Nishant Shah. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581.

Sivasothy Tharsi. 2025. Self-harm detection. https://huggingface.co/datasets/sivasothy-Tharsi/self-harm-detection.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte Peluso. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Yuyou Zhang, Miao Li, William Han, Yihang Yao, Zhepeng Cen, and Ding Zhao. 2025. Safety is not only about refusal: Reasoning-enhanced fine-tuning for interpretable llm safety.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Cross-layer Directional Stability

Figure 5 shows the full $L \times L$ pairwise cosine similarity matrices between contrastive self-harm directions across all layers, for all four models on both datasets. The block-diagonal structure is remarkably consistent across all plots: directions within the same network phase (early or late) are highly aligned, while directions from opposite phases are nearly orthogonal. Gemma-3-4B (rightmost column) stands out with an abrupt near-zero or negative cosine stripe at layers 5 through 8, particularly on SH-Detection, indicating a sharper directional transition than the other models. Summary statistics are reported in Table 4.

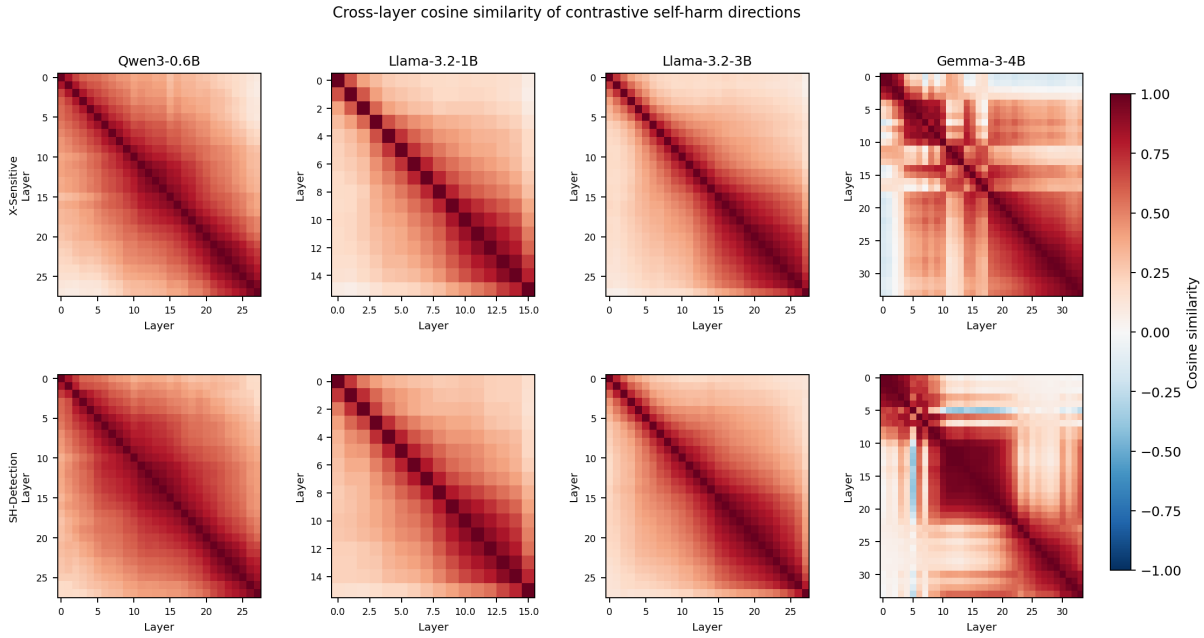

Figure 5: Cross-layer cosine similarity of contrastive self-harm directions ($L \times L$ matrices) for X-Sensitive (top row) and SH-Detection (bottom row). High within-comparison and near-orthogonal cross-comparison cosines confirm the block-diagonal structure across all four models and both corpora.